(19) BUNDESREPUBLIK DEUTSCHLAND

DEUTSCHES PATENT- UND MARKENAMT

(12) **Offenlegungsschrift**

(10) **DE 101 59 262 A 1**

(51) Int. Cl.[7]: C 12 Q 1/68

(21) Aktenzeichen: 101 59 262.0
(22) Anmeldetag: 3. 12. 2001
(43) Offenlegungstag: 18. 6. 2003

(71) Anmelder:
Siemens AG, 80333 München, DE

(72) Erfinder:
Schürmann, Bernd, Prof., 85778 Haimhausen, DE;
Stetter, Martin, Dr., 85667 Oberpframmern, DE

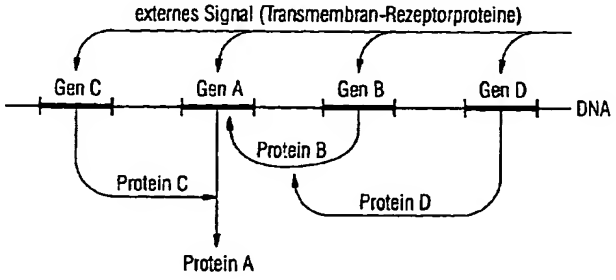
(56) Entgegenhaltungen:
US 62 40 374 B1
Internetdokument, Adresse www.genomatica.com/science tech future.htm (gutachtlich) (recherchiert am 25.07.2002);
Instrinsic noise in gene regulatory networks, TATTAI, M. & VAN OUDENAARDEN, A., Proc. Natl. Acad. Sci. USA (17.07.2001) 98 (15) 8614-8619;
Internetdokument, Adresse www.biosource.com/content/techCornerContent/theSource/SourceIssue6-BSI1.pdf (Frühjahr 2001), Band 6, S. 1 u. 7 (recherchiert am 26.07.2002;

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen

Prüfungsantrag gem. § 44 PatG ist gestellt

(54) Identifizieren pharmazeutischer Targets

(57) Zum Identifizieren pharmazeutischer Targets wird mindestens eine Korrelation zwischen den Expressionsraten verschiedener Gene einer Zelle durch Auswertung einer Mehrzahl von Gen-Expressionsmustern ermittelt. Dabei werden Korrelationen zweiter oder höherer Ordnung betrachtet. Die Korrelationen lassen auf kausale Beziehungen zwischen verschiedenen Genen und den zugehörigen Proteinen schließen. Daher kann aus den Korrelationen das regulatorische Netzwerk der untersuchten Zelle erschlossen werden. Aus dem solcherart erschlossenen regulatorischen Netzwerk können geeignete Targets identifiziert werden.



DE 101 59 262 A 1

DE 101 59 262 A 1

Beschreibung

[0001] Die menschliche Erbsubstanz (Genom) umfasst schätzungsweise 20 000 bis 80 000 Gene, die den genetischen Code für etwa eine Million Eiweißstoffe (Proteine) beinhalten. In den spezialisierten Körperzellen werden jeweils nur Untermengen aller Gene tatsächlich abgelesen (exprimiert). Die Gesamtheit der dadurch erzeugten Proteine wird als Proteom dieser Zelle bezeichnet. Das Wechselspiel der Proteine untereinander sowie mit der DNA stellt den wichtigsten Teil der Maschinerie dar, die der Entwicklung des menschlichen Körpers aus der befruchteten Eizelle sowie allen Körperfunktionen zugrunde liegt. Aus der Sicht der Informatik stellt die Erbsubstanz damit einen prozeduralen Code für die Struktur und Funktion des menschlichen Körpers dar.

[0002] Viele Krankheiten und Fehlfunktionen des Körpers gehen auf Störungen des funktionellen Netzwerks aus Genom und Proteom zurück. So wirken einige Medikamente als Agonisten bzw. Antagonisten spezifischer Zielproteine, d. h. sie verstärken oder schwächen die Funktion eines Proteins mit dem Ziel, das aus Proteom und Genom gebildete regulatorische Netzwerk zurück in einen normalen Funktionsmodus zu bringen. Diese Zielproteine (Targets) werden bislang nach heuristischen Prinzipien aus biochemischen Überlegungen abgeleitet. Oft ist dabei unklar, ob die Fehlfunktion eines Proteins tatsächlich die Krankheitsursache oder nur eines der Symptome einer versteckten Fehlregulation an anderer Stelle des Netzwerks darstellt.

[0003] Für die Entwicklung verbesserter Therapien ist daher ein quantitatives Verständnis des Wechselspiels zwischen Genom und Proteom erforderlich.

[0004] Aufgabe der Erfindung ist es, das Identifizieren von Proteinen, die sich als Target medikamentöser Behandlung genetisch bedingter Krankheiten oder Störungen eignen, zu verbessern.

[0005] Diese Aufgabe wird durch die Erfindungen gemäß den unabhängigen Ansprüchen gelöst. Vorteilhafte Weiterbildungen der Erfindungen sind in den Unteransprüchen gekennzeichnet.

[0006] Zum Identifizieren pharmazeutischer Targets wird mindestens eine Abhängigkeit oder statistische Korrelation zwischen den Expressionsraten verschiedener Gene einer Zelle durch Auswertung einer Mehrzahl von Gen-Expressionsmustern ermittelt. Dabei werden u. a. Korrelationen zweiter oder höherer Ordnung betrachtet. Die Abhängigkeiten lassen auf kausale Beziehungen zwischen verschiedenen Genen und den zugehörigen Proteinen schließen. Daher kann aus den Abhängigkeiten das regulatorische Netzwerk der untersuchten Zelle erschlossen werden.

[0007] So lassen sich Gene identifizieren, die am wahrscheinlichsten regulatorische Kaskaden initiieren, oder die für komplexe Änderungen in den Expressionsmustern, beispielsweise bei Vorliegen einer genetisch bedingten Erkrankung, verantwortlich sind.

[0008] Das erfindungsgemäße Verfahren erlaubt dadurch die Identifizierung von Targets auf systematischer Basis. Dies geschieht durch statistische Modellierung des regulatorischen genetischen Netzwerks mit einem strukturlernenden kausalen Netz auf der Basis von Gen-Expressionsmustern.

[0009] Das beschriebene Verfahren ist nicht auf zeitliche Informationen angewiesen und damit auf eine breite Basis von Gen-Expressionsmessungen anwendbar.

[0010] Das beschriebene Verfahren wird üblicherweise mit Hilfe eines Computers durchgeführt.

[0011] Die Erfindung ist besonders geeignet, High Throughput Drug Discovery Verfahren in der Biotechnologie zu ergänzen. Eine weitere Anwendung der Erfindung findet sich im Bereich der Unterstützung von Tumordiagnostik und Tumorbehandlung. Untersucht werden können sowohl regulatorische Zusammenhänge im menschlichen Körper als auch in jedem anderen Lebewesen, sei es Tier oder Pflanze, Bakterium oder eine sonstige Zelle.

[0012] Die einzelnen Messungen der Gen-Expressionsmuster werden dabei als unabhängig voneinander angesehen. Sie stellen Zufallswerte dar, die von einer unbekannten hochdimensionalen Wahrscheinlichkeitsverteilung erzeugt wurden. Die vollständige Charakterisierung der statistischen Struktur bzw. der Korrelationen der Gen-Expressionsraten anhand der gemessenen Expressionsmuster ist gleichbedeutend mit der Schätzung der zusammengesetzten, hochdimensionalen Wahrscheinlichkeitsverteilung für diese Muster. Beinhaltet eine Messung die Bestimmung der Expression von 5000 Genen, so ist eine 5000-dimensionale Wahrscheinlichkeitsdichte zu schätzen, was in voller Allgemeinheit große Schwierigkeiten bereitet.

[0013] Kausale Netze nehmen an, dass in den Daten bedingte Unabhängigkeiten existieren. Eine bedingte Unabhängigkeit liegt dann vor, wenn zwei Zufallsvariablen unter der Bedingung voneinander unabhängig sind, dass alle anderen Zufallsvariablen konstant gehalten werden, dass also Korrelationen höherer Ordnung über eine mehrstufige Rückkopplungsschleife zwischen den beiden Zufallsvariablen vernachlässigt werden. Die volle Wahrscheinlichkeitsdichte kann dann durch ein Produkt von niedriger dimensionierten Wahrscheinlichkeitsdichten ersetzt werden.

[0014] Eine besonders effiziente Möglichkeit, die Korrelationen bzw. Abhängigkeiten zwischen den einzelnen Zufallsvariablen, also den Expressionsraten, der hochdimensionalen Wahrscheinlichkeitsverteilung zu erschließen, besteht darin, dass zunächst eine Menge von unabhängigen Zufallsvariablen angenommen wird.

[0015] Sukzessiv wird jeweils diejenige Korrelation dem Netzwerk hinzugefügt, die den Fehler des Netzes für die Erklärung neuer Daten (Generalisierungsfehler) am meisten herabsetzt. Das heißt, es werden diejenigen Korrelationen angenommen, bei denen die tatsächlich gemessenen Gen-Expressionsmuster die höchste Wahrscheinlichkeit unter allen denkbaren Wahrscheinlichkeitsverteilungen aufweisen. Dies wird fortgesetzt, bis sich der Generalisierungsfehler nur noch innerhalb einer vorgegebenen Schwelle verringern lässt.

[0016] Die bevorzugte, einfachste Ausführungsform der Suchstrategien für die Korrelationen erfolgt mit Hilfe der folgenden Schritte:

- zunächst wird diejenige alleinige Kante gesucht, die den Generalisierungsfehler minimiert, sozusagen die beste erste Kante.
- anschließend wird die beste zweite Kante gesucht.
- usw., bis sich der Generalisierungsfehler nicht mehr sinnvoll verbessern lässt.

[0017] Auf diese Weise können sowohl die Korrelationen zwischen den Zufallsvariablen (Expressionsraten) erschlossen werden als auch die Form der hochdimensionalen Wahrscheinlichkeitsverteilung, letztere zumindest qualitativ. Das Erschließen der Korrelationen zwischen den Zufallsvariablen mit der Möglichkeit, diese Korrelationen mit Hilfe von mindestens teilweise gerichteten Graphen darzustellen, wird als Strukturlernen bezeichnet, da hierbei die Struktur des regulatorischen Netzwerks gelernt wird.

[0018] Beim sukzessiven Ergänzen von Korrelationen kann auf vorhandenes Wissen über regulatorische Zusammenhänge zurückgegriffen werden. Auf diese Weise kann das Erschließen der regulatorischen Zusammenhänge weiter beschleunigt und präzisiert werden.

[0019] Dieser insbesondere für hochdimensionale Daten sehr zeitaufwändige Algorithmus lässt sich durch schnelle, fast-optimale Suchstrategien für wichtige Abhängigkeiten entscheidend beschleunigen. Ein bekannter Algorithmus hierfür ist der Greedy-Algorithmus (T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein: "Introduction to Algorithms", 2nd edition McGraw-Hill Columbus, OH (2001)).

[0020] Durch eine künstliche Veränderung einzelner Gen-Expressionsraten kann mit Hilfe der aus den bisher vorliegenden Daten berechneten Struktur des regulatorischen Netzwerks, bzw. der hochdimensionalen Wahrscheinlichkeitsverteilung, das am wahrscheinlichsten resultierende Gen-Expressionsmuster vorhergesagt werden. Dieses kann mit Messungen an erkranktem Gewebe (beispielsweise Tumorgewebe) verglichen werden. Dadurch ist es möglich, die einer krankhaft veränderten zellulären Funktion ursächlich zu Grunde liegende Gen-Gruppe bzw. gegebenenfalls das zu Grunde liegende einzelne Gen zu erkennen und das zugehörige Protein als Target einer medikamentösen Behandlung zu identifizieren.

[0021] Im Folgenden wird die Erfindung anhand von Ausführungsbeispielen näher erläutert, die in den Figuren schematisch dargestellt sind. Gleiche Bezugsziffern in den einzelnen Figuren bezeichnen dabei gleiche Elemente. Im Einzelnen zeigt:

[0022] **Fig. 1** schematisch die regulatorischen Vorgänge, die das Expressionsmuster einer Zelle bestimmen;

[0023] **Fig. 2** einen gerichteten azyklischen Graphen; und

[0024] **Fig. 3** illustriert Möglichkeiten, die Richtung von Kanten in einem gerichteten azyklischen Graphen zu bestimmen.

[0025] **Fig. 1** zeigt die wichtigsten Wechselwirkungen zwischen Genen und Proteinen eines DNA-Abschnitts auf. Die Wechselwirkungen werden als Basis für die Beschreibung des genomischen regulatorischen Netzwerks herangezogen.

[0026] Im oberen Teil der **Fig. 1** ist schematisch angedeutet, wie ein von außen auf die Zelle einwirkendes externes Signal – etwa im Rahmen der interzellulären Kommunikation –, das beispielsweise von einem Transmembran-Rezeptorprotein (z. B. von einem Kalziumkanal) aufgenommen und in geeigneter Weise in das Innere der Zelle übertragen wird, die Produktion der Gene A, B, C und D des DNA-Abschnitts auslöst.

[0027] Es besteht daher prinzipiell auch die Möglichkeit, die Expressionsrate einzelner Gene einer Zelle über die erwähnten Wege von außerhalb der Zellen zu beeinflussen.

[0028] Als ein Gen wird ein nicht notwendigerweise zusammenhängender Abschnitt der DNA bezeichnet, der den genetischen Code für ein Protein oder auch für eine Gruppe von Proteinen enthält.

[0029] Der Produktionsvorgang eines Proteins aus einem Gen, zum Beispiel Protein A ausgehend von Gen A in **Fig. 1**, wird als Expression dieses Gens bezeichnet. Die Übersetzung des DNA-Codes des Gens in die Kette der Aminosäuren des Proteins wird als Translation bezeichnet. Die Rate, mit der Protein A in einem gegebenen Kontext produziert wird, wird seine Expressionsrate genannt.

[0030] Nicht alle Gene werden in einer Zelle exprimiert. Vielmehr unterscheiden sich verschiedene Zelltypen durch ihr Gen-Expressionsmuster. Dies gilt oftmals auch für den Unterschied zwischen kranken und gesunden Zellen.

[0031] Das Expressionsmuster einer Zelle wird durch die in **Fig. 1** schematisch dargestellten regulatorischen Vorgänge bestimmt. Die regulatorischen Vorgänge werden im Wesentlichen von einigen wichtigen Wechselwirkungen zwischen Proteinen und Genen sowie zwischen den Proteinen untereinander bestimmt.

[0032] So kann die Expressionsrate eines Gens A durch das Vorhandensein eines anderen Proteins B reguliert, d. h. erhöht, erniedrigt oder zum Erliegen gebracht werden. In diesem Beispiel wirkt das Protein B regulatorisch auf das Gen A bzw. das Protein A. Zu regulatorischen Proteinen können beispielsweise die Proteinbausteine von Aktivatorkomplexen gerechnet werden. Regulatorische Proteine können sich gleichzeitig auf viele Zielgene auswirken.

[0033] Eine zweite Art der Wechselwirkung besteht in der posttranslationalen Modifikation von Proteinen, d. h. der Modifikation von Proteinen nach der Translation. In der Regel erfolgt die posttranslationale Modifikation eines Proteins im unmittelbaren Anschluss an die Translation, d. h. bevor das Protein in der Zelle wirkt. So werden zum Beispiel viele Proteine von speziellen Enzymen phosphoryliert oder glykolysiert, d. h. das Zielprotein wird durch Anhängen bzw. Abspalten chemischer Gruppen in seinen funktionellen Zustand gebracht oder in einen Zustand versetzt, in dem es nicht mehr wirksam ist. Posttranslationale Modifikation kann also ein Protein gegebenenfalls temporär funktionell einschalten oder ausschalten.

[0034] In **Fig. 1** ist das Protein A ein so genanntes Effektorprotein, d. h. es wirkt innerhalb der Zelle auf andere Substanzen und nicht unmittelbar auf das Genom oder Proteom. In **Fig. 1** modifiziert somit das Protein C im Wege der posttranslationalen Modifikation die Funktion des Effektorproteins A.

[0035] Protein B ist ein regulatorisches Protein, da es die Expressionsrate des Proteins A bestimmt, indem es mit demjenigen DNA-Abschnitt wechselwirkt, der das Gen A enthält. Das Protein D modifiziert somit die Funktion eines regulatorischen Proteins (Protein B) im Wege der posttranslationalen Modifikation.

[0036] Die Nukleinsäuresequenz der menschlichen DNA ist weitestgehend bekannt. Auch die durch die DNA kodierten Gene sind in zunehmendem Maße identifiziert. Nicht ganz so vollständig ist das Wissen über das Proteom, einschließlich der eventuell durch Wechselwirkung zwischen den Proteinen posttranslational modifizierten Proteine. Allerdings erlauben neuere Sequenzierungs- und Hochdurchsatz-Screeningverfahren eine rasche Identifizierung weiterer Gene und Proteine.

[0037] Ein weiterer wichtiger Schritt zur Aufklärung der Expressionsmuster einer Zelle hat sich mit der Entwicklung

von Hochdurchsatz-Hybridisierungstechniken vollzogen. Bei diesen Verfahren wird auf einem so genannten Microarray die Expressionsrate vieler 100 verschiedener Gene gleichzeitig getestet. Mit Hilfe dieser Verfahren ist es möglich, das Gen-Expressionsmuster einer Zelle zu bestimmen.

[0038] Dazu werden in der Regel die in der Zelle synthetisierten mRNA (messenger RNA) bestimmt. Die mRNA ist ein Zwischenprodukt bei der Translation des Gens zum Protein. Die mRNA ist somit eine Vorstufe bei der Bildung des Proteins. Die zu untersuchende Zelle wird zunächst isoliert. Anschließend wird sie aufgeschlossen. Durch geeignete Aufreinigungsschritte wird die mRNA aus der Zelle isoliert. Danach wird die mRNA mittels der reversen Transkriptase in cDNA (complementary DNA) übersetzt. Diese wird mit i. d. R. linearer PCR (polymerase chain reaction) amplifiziert. Die so gewonnene cDNA wird mit Hilfe von geeigneten Microarrays, z. B. DNA-Chips, qualitativ bzw. quantitativ analysiert. Mit modernen Microarrays können die Expressionsraten von 5000 und mehr Genen gleichzeitig vermessen werden.

[0039] Aufgrund dieser verbesserten Techniken steht mittlerweile ein umfangreiches Wissen über das menschliche Genom und Proteom sowie über die Wechselwirkungen zwischen Proteinen und Genen bzw. Proteinen untereinander zur Verfügung.

[0040] Im Folgenden werden zunächst einige für die Aufklärung des regulatorischen Netzwerks benötigte mathematische Begriffe eingeführt.

[0041] Die aus den gemessenen Gen-Expressionsmustern bestimmten Expressionsraten der einzelnen Gene sind die im Folgenden zu betrachtenden Zufallsvariablen. Für Gen i wird die die Expressionsrate repräsentierende Zufallsvariable mit X_i bezeichnet. Werte, die sie annehmen kann, werden mit x_i bezeichnet. Mit

$$X := \begin{pmatrix} X_1 \\ \dots \\ X_k \end{pmatrix} = (X_1, \dots, X_k)^T$$

wird der Zufallsvektor bezeichnet, der aus den Expressionsraten aller k Gene besteht. $()^T$ bezeichnet die Transposition.

[0042] Um die Korrelationen zwischen den Expressionsraten bzw. Zufallsvariablen zu ermitteln, werden verschiedene Momente der Zufallsvariablen betrachtet.

[0043] Das erste Moment des Zufallsvektors X, das auch als Erwartungswert E bezeichnet wird, ist definiert durch

$$EX := (\alpha_1, \dots, \alpha_k)^T := (EX_1, \dots, EX_k)^T.$$

[0044] Aufgrund bekannter statistischer Überlegungen wird der Erwartungswert EX_i der Expressionsraten X_i mit Hilfe des arithmetischen Mittels der beobachteten Expressionsraten x_i über n Messungen von Gen-Expressionsmustern geschätzt:

$$E^{(s)}X_i = \frac{1}{n}\sum_{m=1}^{n} x_{im},$$

wobei x_{im} die für das Gen i in der Messung m ermittelte Expressionsrate angibt und der hochgestellte Index (s) anzeigt, dass es sich um einen geschätzten Wert handelt.

[0045] Die zweiten Momente sind definiert durch

$$\alpha_{ij} := E(X_i \cdot X_j).$$

[0046] Wiederum aufgrund bekannter statistischer Überlegungen wird der für das zweite Moment zu berechnende Erwartungswert $E(X_i \cdot X_j)$ mit Hilfe der folgenden Gleichung geschätzt:

$$E^{(s)}(X_i \cdot X_j) = \frac{1}{n}\sum_{m=1}^{n} x_{im} \cdot x_{jm}.$$

[0047] Das zweite zentrale Moment wird auch als Kovarianz bezeichnet. Es ist definiert durch

$$\mathrm{cov}(X_i, X_j) := \mu_{ij} := E([X_i - EX_i] \cdot [X_j - EX_j]).$$

[0048] Es gilt aufgrund der Linearität des Erwartungswerts

$$\mathrm{cov}(X_i, X_j) = \mu_{ij} = E(X_i \cdot X_j) - EX_i \cdot EX_j = \alpha_{ij} - \alpha_i \cdot \alpha_j.$$

[0049] Die Schätzung der Kovarianz erfolgt in bekannter Weise mittels

$$\mathrm{cov}^{(s)}(X_i, X_j) = \frac{1}{n-1}\sum_{m=1}^{n}\left(x_{im} - E^{(s)}X_i\right)\cdot\left(x_{jm} - E^{(s)}X_j\right).$$

[0050] Die μ_{ii} sind gerade die Varianzen der einzelnen Expressionsraten X_i:

$$\sigma_i^2 := \mu_{ii}.$$

[0051] Ihre Schätzung erfolgt in bekannter Weise über

$$\sigma^{(s)}{}_i^2 = \mu^{(s)}{}_{ii} = \frac{1}{n-1}\sum_{m=1}^{n}\left(x_{im} - E^{(s)}X_i\right)^2 .$$

[0052] Die k × k-Matrix

$$\mathrm{cov}(X, X) := E([X - EX] \cdot [X - EX]^T) = E(X \cdot X^T) - EX \cdot EX^T$$

wird als Kovarianzmatrix von X bezeichnet.

[0053] Die Korrelation der Zufallsvariablen X_i und X_j wird häufig mit Hilfe des Korrelationskoeffizienten (zweiter Ordnung) bestimmt. Dieser ist definiert durch

$$\rho_{ij} := \frac{\mathrm{cov}(X_i, X_j)}{\sigma_i \cdot \sigma_j} .$$

[0054] Er liegt zwischen 1 und +1. Er lässt sich unter Verwendung der angegebenen Schätzungen der Kovarianz und der Varianz ebenfalls schätzen. Ein verschwindender Korrelationskoeffizient deutet auf die Abwesenheit regulatorischer Zusammenhänge hin. Ein signifikant von Null verschiedener Korrelationskoeffizient deutet auf eine statistische und damit regulatorische Abhängigkeiten hin.

[0055] Die obigen Definitionen lassen sich auf dritte, vierte und beliebig höhere Momente verallgemeinern. Insbesondere ist das dritte Moment definiert durch

$$\alpha_{ijk} := E(X_i \cdot X_j \cdot X_k).$$

[0056] Das dritte zentrale Moment ist definiert durch

$$\mu_{ijk} := E([X_i - EX_i] \cdot [X_j - EX_j] \cdot [X_k - EX_k]).$$

[0057] Es wird in bekannter Weise geschätzt durch

$$\mu^{(s)}{}_{ijk} = \frac{1}{n-2}\sum_{m=1}^{n}\left(x_{im} - E^{(s)}X_i\right)\cdot\left(x_{jm} - E^{(s)}X_j\right)\cdot\left(x_{km} - E^{(s)}X_k\right) .$$

[0058] Die Korrelation der Zufallsvariablen X_i, X_j und X_k kann ebenfalls mit Hilfe des Korrelationskoeffizienten dritter Ordnung bestimmt werden. Dieser ist definiert durch

$$\rho_{ijk} := \frac{\mu_{ijk}}{\sigma_i \cdot \sigma_j \cdot \sigma_k} .$$

[0059] Er liegt ebenfalls zwischen −1 und +1 und kann in gleicher Weise wie der Korrelationskoeffizient zweiter Ordnung geschätzt werden.

[0060] In einem bevorzugten Ausführungsbeispiel der Erfindung wird das Vorliegen regulatorischer Abhängigkeiten dadurch ermittelt, dass die Korrelationskoeffizienten daraufhin getestet werden, ob sie signifikant von Null abweichen. Statistisch gesprochen wird die Hypothese getestet, ob der Korrelationskoeffizient verschwindet. Dies kann mit Hilfe verschiedener bekannter statistischer Testverfahren durchgeführt werden. Das bevorzugte Verfahren ist beispielsweise in Bronstein-Semendjajew: "Taschenbuch der Mathematik", Verlag Harm Deutsch, 22. Aufl., 1985, S. 693, beschrieben.

[0061] Die geschilderten Verfahren haben allgemein das Ziel, statistische Abhängigkeiten bzw. Unabhängigkeiten aufzuklären und dadurch das Netzwerk der Beeinflussungen aus den Daten zu extrahieren.

[0062] Reguliert das Protein B das Gen A und sind keine anderen regulatorischen Phänomene vorhanden, so äußert sich diese Eigenschaft in einer statistischen Korrelation oder Antikorrelation beider Expressionsraten über verschiedene Messungen hinweg (statistische Abhängigkeit bzw. Korrelation zweiter Ordnung).

[0063] Die Gegenwart eines Metaregulators wie Protein D in **Fig. 1** drückt sich hingegen in einer statistischen Abhängigkeit dritter Ordnung aus, d. h. in einem nicht verschwindenden Korrelationskoeffizienten dritter Ordnung.

[0064] In einer Zelle existieren viele, teilweise noch unbekannte regulatorische Rückkopplungsschleifen, deren Existenz sich in komplexen statistischen Beziehungen zwischen Expressionsraten ausdrückt.

[0065] Korrelationen werden oft durch gerichtete Graphen zwischen Zufallsvariablen dargestellt (siehe z. B. David Edwards: "Introduction to Graphical Modelling", Springer Texts in Statistics, Springer Verlag, 1995). Derartige Modelle werden daher auch als graphische Modelle bezeichnet.

[0066] Die hochdimensionale Wahrscheinlichkeitsverteilung für die Zufallsvariable

$$X := \begin{pmatrix} X_1 \\ \dots \\ X_k \end{pmatrix} = (X_1, \dots, X_k)^T$$

kann mit Hilfe eines Netzwerks oder Graphen G dargestellt werden, wie er in **Fig. 2** für ein einfaches Beispiel gezeigt ist.

Die Knoten **1**, **2** und **3** entsprechen dabei Zufallsvariablen X_1', X_2 und X_3. Im Rahmen der statistischen Modellierung regulatorischer Zusammenhänge im Genom werden die Zufallsvariablen mit den Expressionsraten identifiziert.

[0067] Im Graphen G gem. **Fig. 2** werden Abhängigkeiten durch gerichtete Kanten dargestellt. Dabei wird die Abhängigkeit der Zufallsvariable X_2 von der Zufallsvariable X_1 durch eine gerichtete Kante **12** vom Knoten **1** zum Knoten **2** dargestellt. Die Abhängigkeit der Zufallsvariable X_3 von der Zufallsvariable X_2 wird durch eine gerichtete Kante **14** von Knoten **2** zum Knoten **3** dargestellt.

[0068] Wird eine Korrelation zweiter Ordnung festgestellt, so wird dies durch eine Kante zwischen zwei Knoten, d. h. zwischen zwei Zufallsvariablen, im Graphen angezeigt. Im Allgemeinen ist es nicht möglich, die Richtung dieser Kante zu ermitteln, d. h. welche der beiden Zufallsvariablen ursächlich für die andere ist. Beobachtet wird lediglich das gleichzeitige Auftreten. Damit kann im allgemeinen auch nicht ermittelt werden, welches von den beiden beteiligten Genen bzw. Proteinen das andere reguliert.

[0069] In bestimmten Fällen kann jedoch die Richtung einer Kante ermittelt werden. **Fig. 3A** zeigt einen solchen Fall. Gezeigt sind drei Knoten **1**, **2** und **3**. Zwischen diesen drei Knoten sind zwei Kanten eingezeichnet, und zwar die Kante **20** zwischen den Knoten **1** und **3** sowie die Kante **22** zwischen den Knoten **2** und **3**. Beide Kanten sind in Richtung auf den Knoten **3** gerichtet. In der Graphentheorie wird ein solcher Fall allgemein als "collider" bezeichnet. Statistisch wird man in einer solchen Konstellation eine Korrelation zweiter Ordnung zwischen den Knoten **1** und **3**, also den zugehörigen Zufallsvariablen, ermitteln, sowie eine weitere Korrelation zweiter Ordnung zwischen den Knoten **2** und **3**. Man wird jedoch keine Korrelationen dritter Ordnung feststellen, da beispielsweise die Zufallsvariablen **1** und **3** sich gegenseitig beeinflussen, ohne jedoch einen Einfluss auf die Zufallsvariable **2** zu haben.

[0070] Übersetzt in die Sprache der regulatorischen Wechselwirkungen zwischen Genen bzw. Proteinen zeigt der Graph gem. **Fig. 3A**, dass das Gen **3** durch Gen bzw. Protein **1** und **2** reguliert wird, jedoch nicht umgekehrt. Wird beispielsweise Gen **1** exprimiert, so wird nach dem Modell gem. **Fig. 3A** auch Gen **3** exprimiert. Dies impliziert jedoch nicht, dass auch Gen **2** exprimiert wird. Werden zwei Korrelationen zweiter Ordnung gefunden, eine zwischen Knoten **1** und Knoten **3** und die andere zwischen Knoten **2** und Knoten **3**, so können die Kanten nicht anders gerichtet sein, da sich sonst eine Korrelation dritter Ordnung zeigen würde (vergleiche **Fig. 3B**).

[0071] Anders verhält es sich im Falle von **Fig. 3B**. **Fig. 3B** zeigt Graphen, die im wesentlichen dem Graph gem. **Fig. 3A** entsprechen und auch in gleicher Weise zu lesen sind. Lediglich die Kanten und ihre Richtungen sind variiert. Alle in **Fig. 3B** gezeigten Graphen weisen ausschließlich eine Korrelation dritter Ordnung zwischen den Knoten **1**, **2** und **3** auf und sind auf der Basis der Korrelationsanalyse nicht unterscheidbar.

[0072] Im Allgemeinen ist es sehr schwierig, auf der Basis von Gen-Expressionsmustern posttranslationale Modifikationen zu erschließen. Allerdings geben Korrelationen dritter Ordnung zumindest einen Hinweis auf solche posttranslationalen Modifikationen.

[0073] Im Folgenden wird das Erkennen des zu einem regulatorischen Netzwerk gehörenden Graphen näher erläutert.

[0074] Die gemeinsame Wahrscheinlichkeitsverteilung der Zufallsvariablen X_1, X_2 und X_3 aus **Fig. 2** kann stets durch ein Produkt bedingter Wahrscheinlichkeiten ausgedrückt werden:

$$P(X_1, X_2, X_3) = P(X_3|X_2, X_1) \cdot P(X_2|X_1) \cdot P(X_1).$$

[0075] Im Graphen G gem. **Fig. 2** werden die bedingten Wahrscheinlichkeiten der rechten Seite durch gerichtete Kanten dargestellt. Dabei wird die bedingte Wahrscheinlichkeit $P(X_2|X_1)$ durch eine gerichtete Kante **12** vom Knoten **1** zum Knoten **2** dargestellt. Die bedingte Wahrscheinlichkeit $P(X_3|X_2, X_1)$ wird durch eine gerichtete Kante **14** von Knoten **2** zum Knoten **3** dargestellt. Derartige Graphen G werden als gerichtete azyklische Graphen (DAG, directed acyclic graph) bezeichnet. Die Graphen G heißen azyklisch, da es in dem betrachteten mathematischen Modell niemals eine zyklische Graphenkonfiguration geben wird, bei der beispielsweise in **Fig. 2** auch noch eine gerichtete Kante vom Knoten **3** zum Knoten **1** verläuft, die einen Kreis schließen würde.

[0076] Bei der bedingten Wahrscheinlichkeit $P(X_3|X_2, X_1)$ stellen die Zufallsvariablen X_1 und X_2 die so genannten Eltern (Pa, parents) der Zufallsvariablen X_3 dar, d. h.

$$Pa(X_3) = \{X_1, X_2\}.$$

[0077] Allgemeinen kann daher eine hochdimensionale Wahrscheinlichkeitsverteilung der Variablen X_i geschrieben werden als

$$P(X_1,\ldots,X_k) = \prod_{i=1}^{k} P(X_i|Pa(X_i)).$$

[0078] Dabei ist mit $Pa(X_i)$ die Menge der Eltern der Variablen X_i bezeichnet.

[0079] Statistische Unabhängigkeiten können in einem solchen Graphen G durch betrachten der Eltern einer Zufallsvariablen bestimmt werden.

[0080] Die Struktur eines solchen Graphen G wird durch Vergleich mit gewonnenen Daten, im vorliegenden Fall den gemessenen Gen-Expressionsmustern, bestimmt. Das statistische Problem kann daher in der folgenden Weise formuliert werden: ausgehend von einem Datensatz

$$D = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_k^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_k^{(2)} \\ \vdots & \vdots & & \vdots \\ x_1^{(n)} & x_2^{(n)} & \dots & x_k^{(n)} \end{pmatrix}$$

von n Realisierungen der Zufallsvariablen $(X_1, \dots, X_k)$ wird derjenige Graph G gesucht, der den Datensatz D am besten wiedergibt.

[0081] Es gibt im wesentlichen zwei Wege, die Struktur eines Graphen G aus den Daten D zu erschließen: Die so genannte "constrained based method" (R. Hofmann: "Lernen der Struktur nichtlinearer Abhängigkeiten mit graphischen Modellen", dissertation.de Berlin, 2000) und die so genannte "score based method" (R. Hofmann: "Lernen der Struktur nichtlinearer Abhängigkeiten mit graphischen Modellen", dissertation.de Berlin, 2000), die zur Ausführung der Erfindung bevorzugt wird.

[0082] Die "constrained based method" versucht, statistische Abhängigkeiten bzw. Unabhängigkeiten aus den Daten zu erschließen, ähnlich wie es weiter oben im Zusammenhang mit der Schätzung von Korrelationskoeffizienten geschildert wurde.

[0083] Die "score based method" sucht den Raum der möglichen Graphen ab und bewertet die Übereinstimmung zwischen den Graphen und den Daten mit Hilfe einer Bewertungsfunktion. Das Modell mit dem besten Wert der Bewertungsfunktion wird ausgewählt. Mögliche Bewertungsfunktionen sind das Bayes-Maß (D. Heckerman: "A Bayesian Approach to learning causal networks", Tech Report MSR-TR-95-04, Microsoft Research 1995), die MDL-Metrik (s. u.) oder die BIC-Bewertungsfunktion (G. Schwarz: "Estimating the dimension of a model", The Annals of Statistics 6(2): 461–464 (1978)).

[0084] Die bevorzugte Bewertungsfunktion ist die MDL-Metrik. MDL steht für "minimum description length". Diese Bewertungsfunktion hat zum Ziel, die Daten durch ein Netzwerk bzw. einen Graphen G möglichst genau mit möglichst wenig Kanten zu beschreiben. Die verwendete Bewertungsfunktion lautet:

$$L(G,D) = \log P(G) - n \cdot H(G,D) - \frac{1}{2} K \cdot \log n \ .$$

[0085] Dabei ist logP(G) die a-priori-Wahrscheinlichkeit (im Sinne einer Bayes-Bewertung), den Grafen G vorzufinden. logP(G) wird für alle Graphen G gleich angenommen. Es kann daher bei der Maximierung von L außer Betracht bleiben.

[0086] n ist die Anzahl der zur Verfügung stehenden, gemessenen Datensätze.

$$H(G,D) = \sum_{i=1}^{k} \sum_{e=1}^{E_i} \sum_{l=1}^{r_i} \sum_{j=1}^{q_{ei}} -\frac{N_{ilej}}{n} \log \frac{N_{ilej}}{N_{iej}}$$

gibt die bedingte Entropie des Graphen G in Anbetracht der Daten D wieder.

[0087] Dabei ist k, wie oben erwähnt, die Anzahl der Zufallsvariablen X_i bzw. die Anzahl der Knoten i. D. h. es wird über alle Knoten summiert.

[0088] E_i ist die Anzahl der unmittelbaren Eltern des Knotens i, d. h. die Anzahl der zum Knoten i hin gerichteten Kanten. D. h. es wird zusätzlich über alle zum Knoten i hin gerichteten Kanten summiert.

[0089] r_i ist die Anzahl der möglichen (diskreten bzw. diskretisierten) Werte x_i, die die Zufallsvariable X_i annehmen kann, die also der Knoten i annehmen kann. D. h. es wird über alle möglichen Werte der Zufallsvariablen X_i bzw. des Knotens i summiert.

[0090] q_{ei} ist die Anzahl der möglichen (diskreten bzw. diskretisierten) Werte X_{ei}, die der unmittelbare Elternknoten e des Knotens i, d. h. die Zufallsvariable X_{ei} annehmen kann. D. h. es wird zusätzlich über alle möglichen Werte der Zufallsvariablen X_{ei} bzw. des Knotens e summiert.

[0091] N_{ilej} ist die Anzahl der Datensätze in denen der Knoten i den Wert x_l hat und der unmittelbare Elternknoten e den Wert x_j hat, gezählt über alle n Datensätze. D. h. es wird die Kante zwischen den Knoten i und e betrachtet und gezählt, wie oft bei den gemessenen Datensätzen die zugehörigen Werte x_l und x_j auftraten. Hier fließen die gemessenen Daten ein.

[0092] Schließlich ist die Normierung

$$N_{iej} = \sum_{l=1}^{r_i} N_{ilej} \ ,$$

d. h. es wird über alle Werte summiert, die der Knoten i annehmen kann.

[0093] Die Entropie ist ein nicht-negatives Maß der Unsicherheit, das maximal ist, wenn die Unsicherheit maximal ist, und das verschwindet, wenn vollständiges Wissen vorliegt.

[0094] K ist gegeben durch:

$$K = \sum_{i=1}^{k} \sum_{e=1}^{E_i} q_{ei} \cdot (r_i - 1) \ .$$

[0095] Vernachlässigt man den Term " 1" in der Klammer, so erkennt man in K die Anzahl aller Kombinationen von Werten, summiert über alle Kanten. Ist die Anzahl der Kanten in einem Graphen G klein, so ist in der Regel auch K klein, weshalb L entsprechend größer ist. Dieser letzte Term der rechten Seite erhöht somit den Wert von L für Graphen mit wenigen Kanten, er bevorzugt somit einfache Graphen. Er wird auch Evidenz genannt.

[0096] Die Bewertungsfunktion L entspricht in etwa dem Logarithmus der Bayes-Wahrscheinlichkeit für den Graphen G, wenn die Daten D beobachtet wurden. Sie entspricht damit in etwa der Likelihood des Graphen G. L wird maximiert, d. h. es wird derjenige Graph G gesucht, der für die gegebenen Daten D die Funktion L maximiert.

[0097] Eine besonders effiziente Möglichkeit, die Kanten des Graphen G zu finden, besteht darin, dass zunächst eine Menge von unabhängigen Zufallsvariablen angenommen wird. Sukzessiv wird jeweils diejenige Kante dem Netzwerk hinzugefügt, die die Funktion L am meisten herabsetzt. Dies wird fortgesetzt, bis Minimum von L erreicht ist.

[0098] Wie bereits erwähnt, lässt sich dies in einer bevorzugten, einfachen Ausführungsart mit Hilfe der folgenden Schritte durchführen:

- zunächst wird diejenige alleinige Kante gesucht, die L minimiert, sozusagen die beste erste Kante.
- anschließend wird die beste zweite Kante gesucht, d. h. diejenige zweite Kante, die zusätzlich zur bereits vorhandenen ersten Kante L am weitestgehenden minimiert.
- usw., bis sich L nicht mehr weiter minimieren lässt.

[0099] Dieser insbesondere für hochdimensionale Daten sehr zeitaufwändige Algorithmus lässt sich durch schnelle, fast-optimale Suchstrategien für wichtige Abhängigkeiten entscheidend beschleunigen. Ein bekannter Algorithmus hierfür ist der bereits erwähnte Greedy-Algorithmus.

[0100] Um nicht nur lokale Maxima der Graphenstruktur zu finden, können bekannte Algorithmen wie simulated annealing oder genetische Algorithmen mit den bereits geschilderten Algorithmen kombiniert zur Suche des optimalen Graphen eingesetzt werden.

[0101] Aus dem solcherart erschlossenen regulatorischen Netzwerk können geeignete Targets identifiziert werden. So erkennt man in **Fig. 1** beispielsweise, dass zur Beeinflussung der Konzentration oder Wirksamkeit des Effektorproteins A sowohl das Gen A selbst als auch die Gene B, C und D als Target dienen können.

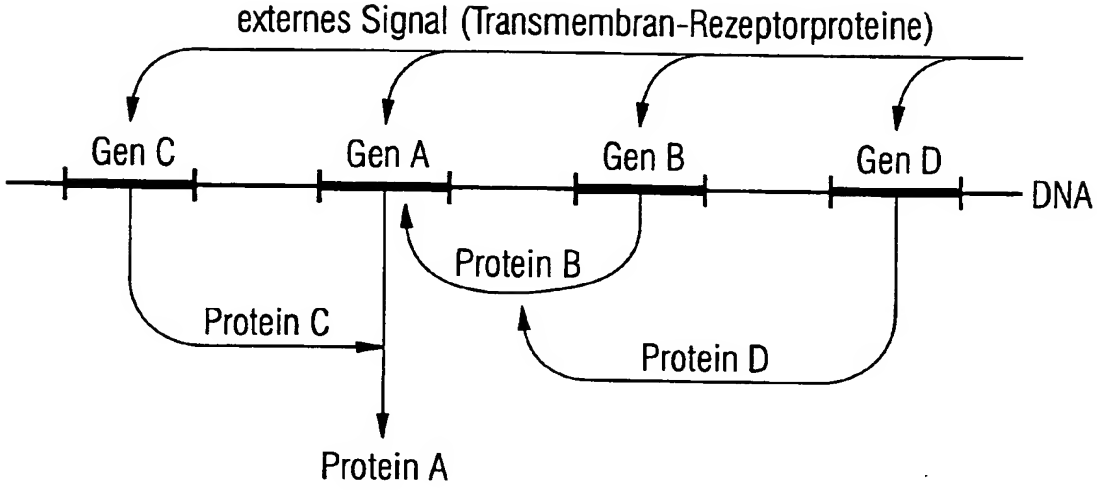
Patentansprüche

1. Verfahren zum Identifizieren pharmazeutischer Targets mit folgenden Schritten:
   a) Eine Mehrzahl von Gen-Expressionsmustern einer Zelle wird bestimmt, wobei jeweils die Expressionsrate der Gene der Zelle bestimmt wird.
   b) Mindestens eine Abhängigkeit zwischen den Expressionsraten der Gene der Zelle wird bestimmt.
   c) Aus mindestens einen Abhängigkeit wird das regulatorische Netzwerk der untersuchten Zelle erschlossen.
2. Verfahren nach dem vorhergehenden Anspruch, dadurch gekennzeichnet, dass angenommen wird, dass nicht alle Expressionsraten der Gene der Zelle voneinander abhängig sind.
3. Verfahren nach dem vorhergehenden Anspruch, dadurch gekennzeichnet,
   dass zunächst eine Menge von unabhängigen Gen-Expressionsraten angenommen wird;
   dass sukzessiv jeweils diejenige Abhängigkeit hinzugefügt wird, die den Fehler für die Erklärung der Gen-Expressionsmuster am meisten herabsetzt.
4. Verfahren nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet, dass die Abhängigkeiten mit Hilfe von Methoden der Graphen-Theorie ermittelt werden.
5. Verfahren nach einem der vorhergehenden Ansprüche, dadurch gekennzeichnet,
   dass die Expressionsrate mindestens eines Gens der Zelle künstlich verändert wird;
   dass mindestens ein Gen-Expressionsmuster der solcherart veränderten Zelle bestimmt wird; und
   dass das bestimmte Gen-Expressionsmuster mit mindestens einem berechneten Gen-Expressionsmuster verglichen wird, das auf der Basis der mindestens einen künstlich veränderten Gen-Expressionsrate berechnet wurde.
6. Anordnung zum Identifizieren pharmazeutischer Targets
   d) mit Mitteln zum Bestimmen einer Mehrzahl von Gen-Expressionsmustern einer Zelle, wobei jeweils die Expressionsrate der Gene der Zelle bestimmt wird;
   e) mit Mitteln zum Bestimmen mindestens einer Korrelation zwischen den Expressionsraten der Gene der Zelle;
   f) mit Mitteln zum Erschließen des regulatorischen Netzwerks der untersuchten Zelle aus den bestimmten Korrelationen.

Hierzu 2 Seite(n) Zeichnungen

Nummer: DE 101 59 262 A1
Int. Cl.[7]: C 12 Q 1/68
Offenlegungstag: 18. Juni 2003

FIG 1



Nummer: DE 101 59 262 A1
Int. Cl.[7]: C 12 Q 1/68
Offenlegungstag: 18. Juni 2003

FIG 2

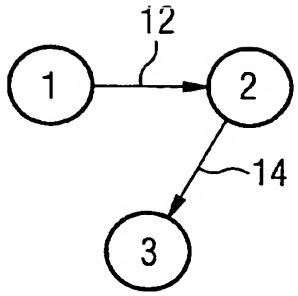


FIG 3A

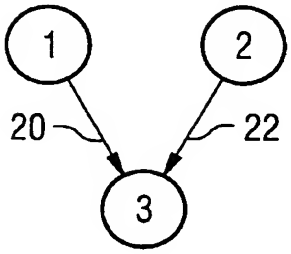


FIG 3B